PHANTOMWIKI: ON-DEMAND DATASETS FOR REASONING AND RETRIEVAL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality benchmarks are essential for evaluating reasoning and retrieval capabilities of large language models (LLMs). However, curating datasets for this purpose is not a permanent solution as they are prone to data leakage and inflated performance results. To address these challenges, we propose *PhantomWiki*: a pipeline to generate unique, factually consistent document corpora with diverse question-answer pairs. Unlike prior work, *PhantomWiki* is neither a fixed dataset, nor is it based on any existing data. Instead, a new *PhantomWiki* instance is generated on demand for each evaluation. We vary the question difficulty and corpus size to disentangle reasoning and retrieval capabilities, respectively, and find that *PhantomWiki* datasets are surprisingly challenging for frontier LLMs. Thus, we contribute a scalable and data leakage-resistant framework for disentangled evaluation of reasoning, retrieval, and tool-use abilities.

1 INTRODUCTION

Designing agents that can perform complex reasoning while interfacing with a large-scale, dynamic corpus—like Wikipedia—is a long-standing goal in the field of natural language processing (Feldman & El-Yaniv, 2019; Min et al., 2019). Such a goal may be within reach given the impressive capabilities of recent language models, which are all trained on internet-scale data. For example, the ability of LLMs to solve math problems on GSM8K (Cobbe et al., 2021) and mathematical olympiads AlphaProof & AlphaGeometry (2024) could bode well for agents to answer highly quantitative questions. On benchmarks like DROP (Dua et al., 2019) and MMLU (Hendrycks et al., 2020), these LLMs demonstrate advanced reading comprehension and general reasoning capabilities, both necessary for intelligent agents. When augmented with retrievers (Muennighoff et al., 2022) and tools (Patil et al., 2023), LLMs seem to already possess a strong ability for accessing external datastores and knowledge bases.

However, it is unclear to what extent these models rely on their internal knowledge, which can easily become outdated, versus their reasoning and retrieval abilities. Consider the example, "*What is the date of birth of Wolfgang Amadeus Mozart?*". Since this fact is contained within LLMs' pre-training data, asking LLMs this question cannot provide reliable insight on whether the answer was deduced, retrieved or recalled. At the same time, existing approaches that perturb Wikipedia facts (Cohen et al., 2024; Meng et al., 2022; Elazar et al., 2021) to construct new question-answer pairs face challenges of ensuring factual consistency across articles. For example, changing Mozart's date of birth to 2025 also requires modifying Beethoven's article to erase the fact that Beethoven might have met Mozart in 1787!

Reasoning-only benchmarks are not immune to memorization either. On GSM8K, a dataset that contains grade school math problems, Mirzadeh et al. (2024) report that frontier models perform significantly worse with minor or even meaningless alterations to the test data, indicating these models are vulnerable to overfitting at best and exact memorization at worst. To ensure fair comparison, LLMs need to be evaluated in a way that does not depend on any particular dataset instance.

Following this philosophy, we introduce *PhantomWiki*. At the click of a button, *PhantomWiki* generates a fictional universe of characters along with a set of facts. We reflect these facts in a large-scale corpus, mimicking the style of fan-wiki websites. Then we generate question-answer pairs with tunable difficulties, encapsulating the types of multi-hop questions commonly considered in the question-answering (QA) literature.

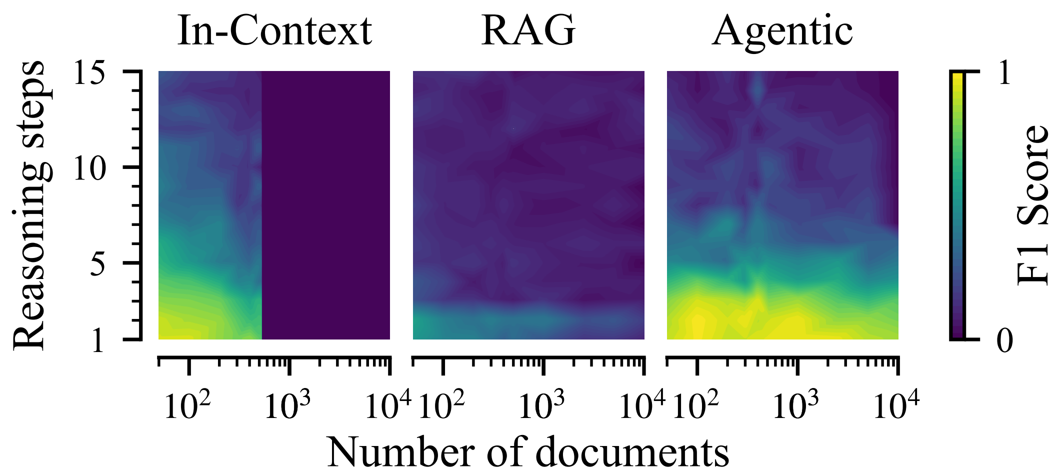

Figure 1: **Evaluating LLM capabilities with *PhantomWiki*.** We tune the reasoning and retrieval difficulty by the number of reasoning steps and documents, respectively.

We design *PhantomWiki* to facilitate testing of different aspects of LLMs. In the first setting, the universe is small enough such that all relevant information can fit within the context. Studies such as (Liu et al., 2024) show that LLMs perform poorly in "needle-in-a-haystack" scenarios, where a small but crucial piece of information is embedded within a long document. By adjusting the total context length—determined by *PhantomWiki* universe size—and the quantity of relevant information required for a given question, *PhantomWiki* provides a reliable benchmark for evaluating LLMs' in-context retrieval capabilities.

In the second setting, when dealing with large-scale corpora, LLMs face inherent limitations in processing all available information within their fixed context window. Instead, they must rely on external retrieval methods to access relevant information (Lewis et al., 2020). This setup allows us to decouple and evaluate two key components: the effectiveness of the retriever in identifying and retrieving the most relevant content, and the LLMs' ability to accurately interpret and utilize the retrieved information.

Last but not least, where LLMs are augmented with external tools, effectively integrating reasoning and tool-use capabilities becomes essential for solving complex tasks. By adjusting the size of the associated text corpus and modulating the reasoning difficulty, *PhantomWiki* serves as a foundation for future research on agents that can seamlessly combine reasoning and tool utilization.

Our evaluation on *PhantomWiki* confirms that the proposed tasks present significant challenges for all of the state-of-the-art LLMs that we used. As we show in Figure 1, we observe consistent performance declines across all scenarios as the universe size grows or the number of reasoning steps increases, indicating heightened difficulty in retrieval and reasoning. By breaking down challenges across different dimensions, *PhantomWiki* enables researchers from various fields to evaluate and refine their methods. Beyond serving as a robust benchmark for LLM performance, *PhantomWiki* provides valuable insights that can guide improvements in retrieval, reasoning, and tool-use capabilities of LLMs for the research community. We will make *PhantomWiki* code available in a public GitHub repository after the anonymous reviewing period.

2 Related Works

Agent benchmarks, such as τ-bench (Yao et al., 2024), ToolWoz (Lattimer et al., 2024), Alfworld (Shridhar et al., 2020) and WebArena (Zhou et al., 2023), focus on tasks where the agent is given a binary reward for successful completion of a task (e.g., booking a flight, making a purchase). In this work, we focuses more on tasks where the agent is rewarded for responding to a question with a factually correct answer. (In Section 3, we concretize what we mean by a "fact" and "correctness".) Zhou et al. (2023, Section 3.2) include a category of information-seeking tasks, however these tasks

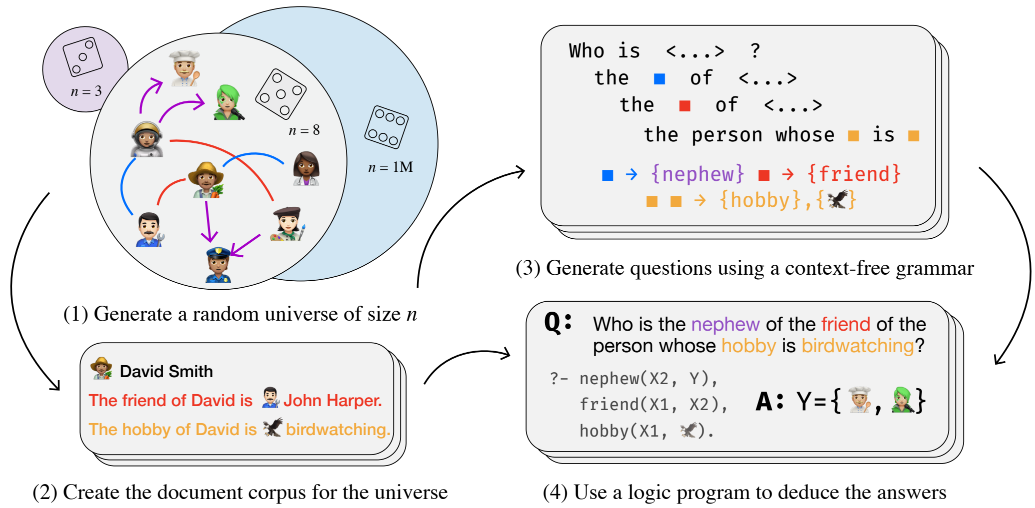

Figure 2: **Overview of the *PhantomWiki* pipeline.**

often require navigation across multiple pages or focus on user-centric content. Yao et al. (2024, Appendix. A) measure task difficulty based on the average success rate of frontier models (e.g., GPT-4). *Our work defines a model-agnostic measure of difficulty, which we show provides more meaningful insight into the reasoning and retrieval aspects of LLMs.*

In the QA domain, existing benchmarks are designed to test whether LLMs are able to reason and use tools. Closer to our work in the space of question-answering agents is the ToolQA benchmark of Zhuang et al. (2023). They introduce a framework to construct question-answer pairs from databases and documents by first generating question templates using LLMs, then filtering for high-quality templates, and finally deriving ground-truth answers by writing corresponding Python programs for each question template. Zhuang et al. (2023, Tab. 1) construct two pure-text datasets: SciREX with 438 documents and 5 question templates, and Agenda with 10k event entries and 10 question templates. *In constrast, PhantomWiki generates instances at much larger scale with 50 question templates and 1 million documents.*

For generating factually-grounded answers, retrieval augmented generation (RAG) has emerged as the predominant paradigm (Lewis et al., 2020; Karpukhin et al., 2020; Guu et al., 2020). However, evaluating RAG systems is notoriously difficult, leading to a flourishing of retrieval benchmarks (Petroni et al., 2020; Saad-Falcon et al., 2023; Jin et al., 2024; Hsia et al., 2024; Mao et al., 2024; Rau et al., 2024). A key pain-point of RAG is handling questions that involve multi-hop reasoning. This motivated Tang & Yang (2024) to design the MultiHop-RAG dataset with synthetically generated questions and Su et al. (2024) to curate a dataset of question-answer pairs that requires intensive reasoning to retrieve relevant documents. *Importantly, none of these benchmarks creates the underlying corpus, a limitation which we bridge in this work.*

Logical reasoning tasks have become central to LLM evaluation and have garnered significant attention in recent time (Zhu et al., 2023). Many existing benchmarks do not disentangle the evaluation on logical reasoning with other abilities such as natural language inference and commonsense reasoning (Sakaguchi et al., 2021; Zellers et al., 2019; Sprague et al., 2023). A line of works focuses on the synthesis of datasets containing a variety of logic reasoning tasks (Tafjord et al., 2020; Saparov & He, 2022; Liu et al., 2020; Han et al., 2022; Weston et al., 2015). Closer to our work, Sinha et al. (2019) construct short stories about individuals related through a family graph and ask questions about their kinship relationships to benchmark the inductive reasoning capabilities. *However, theirs is distinct from our work in that all relevant information for a specific question is centralized in a single article; PhantomWiki requires that the relevant information first be retrieved from a large-scale corpus.*

3 *PhantomWiki* CONSTRUCTION

PhantomWiki is at its core an on-demand random generator of fictional worlds. Similarly to the wiki hosting services popular in film, video games and literature[1], we represent these fictional worlds through Wikipedia-like biographical entries about their characters. We then test the model's retrieval skills and its understanding of the fictional world through an accompanying set of automatically generated question-answer pairs.

3.1 GENERATING A *PhantomWiki* UNIVERSE

The first part of the *PhantomWiki* pipeline generates a random universe of n characters as well as the document corpus describing it, as illustrated in Figure 2, (1-2).

Generating Characters. Each character in a *PhantomWiki* universe is described through its *social relationships* and *personal facts* (Figure 2, (1)). For the social relationships, we first generate family trees, following the family tree generator from Hohenecker & Lukasiewicz (2020). We iteratively pick a person and generate their parent or child based on various constraints[2], until the user-specified universe size of n people is reached. The user can also specify other hyperparameters like the number of trees, their maximal depth, and the maximal number of offspring for each person. In addition to the family trees, we generate a friendship graph using the Erdős–Rényi model (making two people friends with some fixed probability, typically controlled by the desired average number of friendships.)

Generating Facts. Next, we generate personal facts for each person in the *PhantomWiki* universe. Names are assigned during the family generation procedure, with the first name sampled based on the character's gender and the surname based on the family tree, resulting in 15M full names in total[3]. We also add dates of birth in a way that is consistent with the existing family relations, and assign each person a job and a hobby that we uniformly sample from over 300 and 600 options respectively.

Generating Articles. Given all relevant facts for each person, we convert them into articles using pre-defined templates, e.g. "The job of David is a farmer. The hobby of David is birdwatching." (see Figure 2, (2)). This construction conveys the necessary information while keeping the articles short (about 160 tokens on average). While it is possible to extend the article generation process to LLM-based methods (see e.g. Shao et al. 2024), this poses the challenge of guaranteeing factual correctness without additional costs and external supervision. This has been supported by our preliminary experiments on article generation using Llama-3.3-70B, where we observed factual errors in the resulting articles; therefore we do not use LLMs and rely entirely on templates. The articles are the only component of *PhantomWiki* available to the model during its evaluation.

3.2 GENERATING QUESTION-ANSWER PAIRS

In the second half of the *PhantomWiki* pipeline, we generate a set of questions with verifiable answers, as shown in Figure 2, (3-4).

Generating Questions. We implement automatic question generation through a context-free grammar (CFG, Hopcroft et al. 2001) of *question templates*, which we then use to sample complete questions. For example, the question template "*Who is the* <relation> *of* <name>?" can be used to sample the question "*Who is the friend of David?*" (see Figure 2, (3)). The main advantage of using a CFG is that it efficiently and systematically obtains *all* possible compositions of questions for some recursion limit d. For instance, the following subset of our context-free grammar:

$$\begin{aligned} S &\rightarrow \text{Who is } R\ ? \\ R &\rightarrow \text{the } <\textit{relation}> \text{ of } R' \\ R' &\rightarrow R \mid <\textit{name}> \end{aligned}$$

can lead to questions ranging from "*Who is the friend of David?*" to "*Who is the nephew of the friend of the brother of David?*" as d increases. In addition to these nested compositions, our CFG also

[1] For example, see stardewvalley.fandom.com or harrypotter.fandom.com.

[2] For example, the number of offspring of a person has to be smaller than some threshold, parents of the people at the maximal tree level will not be generated, etc.

[3] We use unique names in our experiments, but *PhantomWiki* also supports repeated names.

supports questions about personal attributes (e.g. "*Who is the person whose hobby is birdwatching?*"), aggregation questions ("*How many brothers does David have?*"), and combinations of all three ("*How many friends does the brother of the person whose hobby is birdwatching have?*") (For the full CFG see Appendix B.)

Generating Answers. To ensure that the answers to the sampled questions are verifiably correct, we represent our generated universe in Prolog, a logic programming language (Sterling & Shapiro, 1994). Each Prolog program consists of a set of facts known about the world such as `hobby("David", "birdwatching")`, and a set of rules defining how facts are related to each other, such as `nephew(X, Y) :- sibling(X, A), son(A, Y)`. The Prolog program uses these facts and rules to deduce the exhaustive set of answers to its *queries* (i.e., the CFG-generated questions). For example, a question "*Who is the nephew of the friend of the person whose hobby is birdwatching?*" corresponds to the three-statement Prolog query `?- nephew(X2, Y), friend(X1, X2), hobby(X1, "birdwatching")`, which returns all people satisfying these constraints in the *PhantomWiki* universe (see Figure 2 (4)).

To construct the Prolog queries automatically, we modify the CFG algorithm to generate both the question and query templates in parallel. We note, however, that the queries are separate from the final *PhantomWiki* corpus and question-answer pairs, and the answers returned by the Prolog program should be held out as part of the evaluation procedure.

3.3 *PhantomWiki* Complexity

The goal of *PhantomWiki* is to generate memorization-resistant evaluation datasets that are challenging in both reasoning and retrieval aspects. In this section, we discuss our conceptual and practical design choices that help us achieve this goal.

Universe Space Complexity. To ensure that our evaluation with *PhantomWiki* is memorization and data leakage-resistant, we first show that the space of possible universes is sufficiently large to generate enough unique instances. Observe that the number of possible friendship assignments grows at the rate of $\Theta(2^{n^2})$ (Flajolet & Sedgewick, 2009, Ex. II.5) as the number of individuals n in the universe increases. Similarly, assuming each individual is assigned one fact from each category (job, hobby, etc.), the number of possible fact assignments grows at the rate $\Theta(c^n)$, where c is the total number of choices across the categories. *PhantomWiki* thus samples a corpus from $\Theta(2^{n^2} c^n)$ possible universes, which leads to diverse datasets optimal for data leakage-resistant evaluation. We note that as future work *PhantomWiki* could be extended to increase this diversity, e.g. by adding a temporal dimension of events.

Reasoning Complexity. The CFG enables us to recursively compose templates that lead to complex reasoning questions. Observe that our CFG in Appendix B produces $\Theta(d)$ question templates as the recursion limit d increases. Moreover, we can increase the difficulty of each template by increasing the number of *reasoning steps*. For example, substituting *<relation>* with *nephew* in a template adds two reasoning steps (`nephew(X, Y) :- sibling(X, A), son(A, Y)`), since *PhantomWiki* articles only contain immediate family relationships like *sibling* and *son*. In contrast, substituting *<relation>* with *second cousin* would lead to five reasoning steps. As we will show in Section 4, *PhantomWiki* questions are sufficiently complex to evaluate reasoning capabilities of state-of-the-art LLMs. We further note that *PhantomWiki*'s CFG can be easily extended to support more question types like comparison and multiple-constraint questions.

Retrieval Complexity. To assess a model's retrieval capabilities, we increase the universe size n so that the document corpus exceeds the model's context length—this makes a retriever necessary to answer questions correctly. For state-of-the-art LLMs with a context length of 128K, such as OpenAI's GPT-4o and Meta's Llama-3.3-70B, this corresponds to *PhantomWiki* universes of $n \gtrapprox 1$K. This increases to $n \gtrapprox 3$K for Google's Gemini-1.5-Flash with context length 1M. Further scaling n leads to further increase in retrieval difficulty. In Table 1, we show that *PhantomWiki* is well-suited for generating universes of this size on standard CPU hardware: generating 10-hop questions for size $n = 100$K—well beyond any existing LLM's context length—takes just 6 minutes on 8 Intel Cascade Lake CPU cores. Moreover, we can conveniently generate instances of $n = 1$M, which is on the scale of Wikipedia's corpus of 2 million biographical entries[4].

[4] WikiProject Biography, as of January 30, 2025.

Table 1: Runtimes of generating a *PhantomWiki* instance for different universe sizes n.

n	Runtime
10^2	1 second
10^3	3 seconds
10^4	21 seconds
10^5	6 minutes
10^6	4 hours

Table 2: **F1 scores for various LLMs and prompting techniques.** We report mean ± standard error across 3 dataset generation seeds (except for GPT-4o due to cost constraints), and indicate the highest F1 score for each n in bold. In-Context prompting is infeasible for $n = 5$K as the corpus cannot be fully included in the context.

n	Model	In-Context ZeroShot	In-Context CoT	RAG ZeroShot-RAG	RAG CoT-RAG	Agentic ReAct
50	DeepSeek-R1-32B	42.42 ± 1.69	**52.42 ± 2.64**	16.87 ± 0.98	16.10 ± 1.10	—
50	GPT-4o	27.20	50.66	20.54	14.14	38.70
50	Gemini-1.5-Flash	28.49 ± 1.15	34.61 ± 2.41	19.88 ± 2.05	13.35 ± 0.66	30.92 ± 1.41
50	Llama-3.3-70B	25.64 ± 0.56	48.05 ± 1.95	17.55 ± 2.20	20.01 ± 1.81	35.83 ± 1.00
500	DeepSeek-R1-32B	18.33 ± 2.33	19.65 ± 3.00	12.08 ± 1.07	9.64 ± 0.46	—
500	GPT-4o	16.76	**41.02**	13.56	7.25	37.39
500	Gemini-1.5-Flash	17.39 ± 1.45	25.17 ± 1.77	11.66 ± 0.34	7.94 ± 0.19	26.99 ± 1.84
500	Llama-3.3-70B	11.59 ± 1.19	28.07 ± 0.37	10.89 ± 0.58	11.54 ± 1.05	35.56 ± 0.49
5000	DeepSeek-R1-32B	Max context exceeded		8.29 ± 0.18	7.96 ± 0.38	—
5000	GPT-4o			10.12	6.96	**36.85**
5000	Gemini-1.5-Flash			8.60 ± 0.31	5.26 ± 0.36	23.47 ± 1.53
5000	Llama-3.3-70B			7.57 ± 0.52	8.67 ± 0.18	30.89 ± 2.24

4 Experimental Validation

We evaluate reasoning and retrieval capabilities of several frontier LLMs using *PhantomWiki*, by decomposing their performance over questions of varying difficulty and universes of varying sizes.

4.1 Evaluation Setup

We generate *PhantomWiki* instances with n ranging from 50 to 10K—a universe size for which the total length of articles exceed the LLM context length. For the evaluation, only the articles (not the Prolog database or the generated graphs) will be provided to the LLMs. To ensure that our findings are not tied to any specific *PhantomWiki* instance, we use 3 random dataset seeds for each configuration. Creating *PhantomWiki* instances with different random seeds leads to entirely different combinations of names, relations, and personal facts. In each instance, we generate question templates with maximum recursion depth $d = 20$, for a total of 50 templates. We sample 10 questions for each template, yielding a **total of 500 questions per *PhantomWiki* instance**. As shown in Figures 5 and 6 (Appendix B), these questions have varying difficulty and number of answers. Accordingly, we prompt the LLMs to predict all answers as a comma-separated list and measure correctness with the answer-level F1 score.

4.2 Models and Prompting Techniques

We test both open- and closed-source LLMs, namely OpenAI's GPT-4o (Hurst et al., 2024), Google's Gemini-1.5-Flash (Gemini Team, Google, 2024), and the instruction-tuned version of Meta's Llama-3.3-70B model (Dubey et al., 2024). We also evaluate DeepSeekAI's DeepSeek-R1-32B (Guo et al.,

2025) (distilled with Qwen-2.5-32B (Yang et al., 2024)), which is an open-weights LLM trained on reasoning trace datasets. We prompt each LLM with the following techniques, broadly grouped in three ways:

In-Context Prompting. This technique includes the whole document corpus as part of the prompt. We use this type of prompting in conjunction with two strategies: ZEROSHOT—where the document corpus is immediately followed by the question—and Chain-of-Thought (COT) prompting (Wei et al., 2022), where we additionally include some examples on how step-by-step reasoning could lead to the correct answer. We include these prompts in Appendix C.2, as well as modifications to the ZEROSHOT strategy for DeepSeek-R1-32B.

RAG Prompting. This setting augments generation with a pre-trained neural retriever (Lewis et al., 2020). We implement this by first searching for the 4 most relevant documents to the posed question based on UAE-LARGE-V1 embeddings. Next, we incorporate these retrieved documents into the model's prompt. Finally, we add in the same ZEROSHOT and COT prompts as In-Context Prompting. We document details on our retrieval algorithm in Appendix C.3.

Agentic Prompting. REACT (Yao et al., 2022) is a prompting technique that enables LLMs to interleave reasoning steps with tool interactions, to solve complex tasks. For *PhantomWiki* QA task, the LLMs are provided with keyword-based tools `RetrieveArticle` and `Search` to retrieve relevant documents (see Appendix C.6 for tool details). These settings materialize the limitations of in-context prompting and necessitate the use of advanced RAG prompting and agentic prompting approaches.

In the COT and REACT prompts, we include 10 QA exemplars and hand-written reasoning traces. We choose these exemplars from a dataset instance of size 25 that is not used for evaluation. In REACT, we limit LLMs to interact with the text corpus for up to 50 steps, which is sufficient to answer almost all questions in *PhantomWiki* instances.

We cap all LLM outputs to 4096 tokens and use greedy decoding (temperature $= 0$). For DeepSeek-R1-32B, we use temperature $= 0.6$ and top-p $= 0.95$ in accordance with the evaluation setup in Guo et al. (2025, Sec. 3). We refer the reader to Appendix C for full prompt templates and implementation details.

4.3 DISCUSSION

In Table 2, we report the mean F1 score across various universe sizes, LLMs, and prompting techniques. We first average F1 scores over all questions in a *PhantomWiki* instance, then compute the mean and standard error across the dataset generation seeds.

We first consider the small-universe setting ($n = 50$) in Table 2, which corresponds to roughly 16K tokens for the LLMs we test. In-Context prompting techniques outperform other techniques: COT with GPT-4o attains the highest performance, followed by ZEROSHOT with DeepSeek-R1-32B. Next, we consider the setting of medium universes ($n = 500$). Here the full document corpus can still be included in all LLMs' contexts, but we find that ZEROSHOT performs poorly for all LLMs, and DeepSeek-R1-32B especially struggles. F1 scores of COT for all LLMs degrade as well compared to $n = 50$, but not worse than REACT. Finally, in the setting of large universes ($n = 5000$), none of the LLMs we evaluate can accommodate the full document corpus. in-context prompting techniques are no longer viable, and we must rely on RAG prompting and agentic prompting. RAG prompting attain poor F1 scores because the retriever fails to retrieve documents relevant for answering complex questions. On the other hand, agentic prompting technique shines in comparison to other techniques, indicating that LLMs are better suited to dynamically retrieve documents while reasoning on a question.

5 EVALUATING REASONING

To isolate LLM reasoning capabilities with *PhantomWiki*, we investigate model performance on small universes ($n = 50$) in Figure 3. Note that contexts of all LLMs can fully include small universe document corpora. Each *PhantomWiki* dataset contains questions covering a wide range of difficulty. We evaluate three approaches: in-context prompting, RAG prompting, and agentic prompting. For each we plot the F1 scores as a function of question difficulty, as measured by the number of *reasoning*

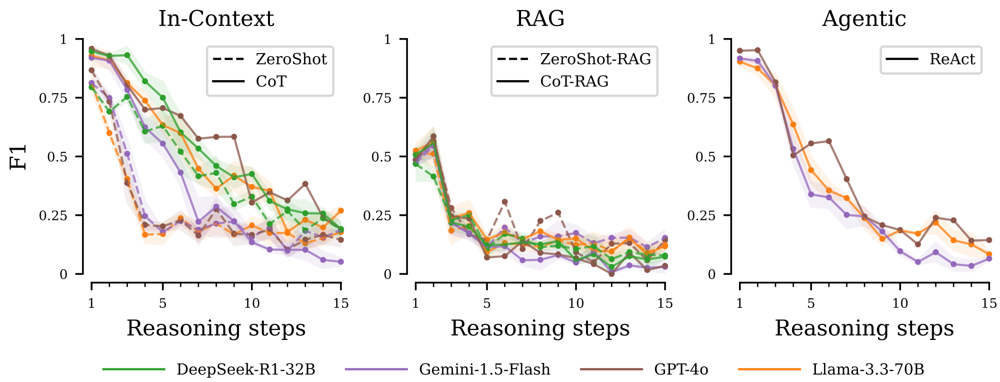

Figure 3: **F1 scores as a function of question difficulty, measured by *reasoning steps*.** We plot LLM performance on universe size $n = 50$, and report F1 scores averaged over 3 generation seeds. Increasing question difficulty in *PhantomWiki* reveals a clear decline across all state-of-the-art LLMs and prompting techniques, showing their struggle with reasoning.

steps necessary to answer the question. As mentioned in Section 3.3, this is determined by the type of question templates and the sampled relationships. For all LLMs and prompting techniques, we verify empirically that **questions with larger reasoning steps are indeed more challenging to answer**. By allowing question difficulty to be adjusted, *PhantomWiki* serves as a foundational benchmark for evaluating reasoning capabilities in language models.

ZeroShot performance declines sharply as the number of reasoning steps increases for all LLMs, except for DeepSeek-R1-32B, which deteriorates more gradually. LLMs perform better with CoT than with ZeroShot, but each additional reasoning step remains increasingly challenging. This suggests that even in the absence of retrieval constraints, LLMs struggle to navigate logical reasoning sequences.

RAG prompting techniques (ZeroShot-RAG and CoT-RAG) stunt reasoning performance across the board—F1 scores are near zero on questions with 5 or more reasoning steps as opposed to 15 steps for in-context prompting. We attribute this to a core problem with RAG prompting: retrieving documents in the initial prompt before starting to answer the question, as opposed to reasoning through the question and retrieving documents dynamically.

We find that RAG prompting techniques can only answer questions that require a single reasoning step, like *Who is the friend of David?*. On the other hand, answering questions that require information from *multiple* reasoning steps is extremely challenging for ZeroShot-RAG and CoT-RAG. To illustrate, consider the question *Who is the nephew of the friend of David?* Answering this question requires retrieving David's document first and then retrieving their friend's document to find the nephew. Since RAG prompting techniques retrieve documents *only once* by matching vector embeddings of questions and documents, they are unlikely to retrieve all necessary documents required to answer such questions.

Finally, the agentic prompting technique ReAct allows LLMs to avoid the steep performance drop as seen in RAG prompting. On given a question, ReAct prompting requires LLMs to retrieve documents dynamically in a conversation and justify why they are relevant. Concretely, before using a tool (`RetrieveArticle` or `Search`) in a conversation turn, the LLM is asked to describe how the tool will help using a "Thought" step (Yao et al., 2022), analogous to the CoT prompting approach. This approach shows promise in answering questions correctly. Even so, ReAct struggles as the question difficulty increases.

Figure 3 thus decomposes LLM performance along the lines of reasoning capabilities. It reveals that all in-context prompting and agentic prompting achieve near-perfect F1 scores on low-difficulty questions. Therefore, the stratification between them in Table 2 can be attributed to varying performance on high difficulty questions. To further isolate the impact of question difficulty, in Figure 7 we plot F1 scores as a function of reasoning steps for questions with only one solution.

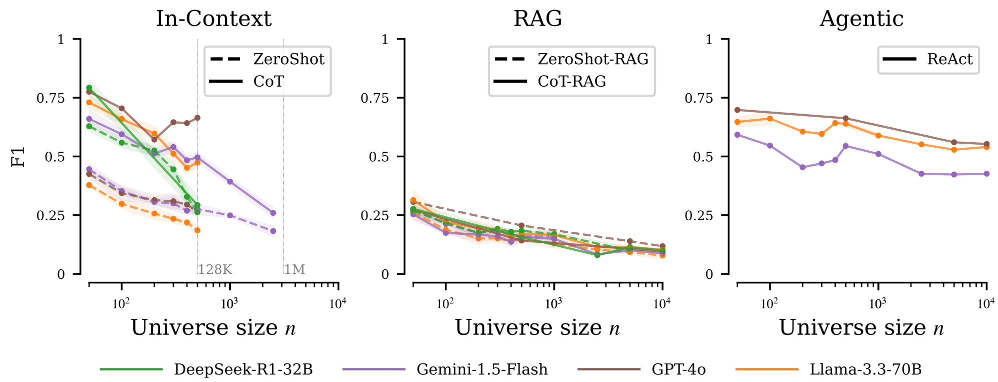

Figure 4: **F1 scores as a function of universe size n.** We evaluate LLM performance on questions with ≤ 10 reasoning steps, and report F1 scores averaged over 3 dataset generation seeds. As we increase universe size in *PhantomWiki*, F1 scores for all LLMs and prompting techniques deteriorate, highlighting that they struggle at retrieving relevant documents.

6 Evaluating Retrieval

Next, to evaluate LLM retrieval capabilities, we use *PhantomWiki* to contrast two settings: (1) small universes where the document corpus can comfortably fit in LLM context, and (2) large universes where the full corpus exceeds context lengths. To this end, we increase the universe size up to $n = 10$K, which corresponds to document corpora well beyond the context lengths of state-of-the-art LLMs, and display the results in Figure 4.

For very small universes, CoT usually outperforms ZeroShot for all LLMs except DeepSeek-R1-32B. However, F1 scores noticeably worsen as more documents are included in models' contexts, with DeepSeek-R1-32B suffering a dramatic performance drop. This analysis regime indicates that **state-of-the-art LLMs struggle at in-context retrieval** for complex question-answering tasks.

At the large universe scale, in-context prompting techniques become nonviable as the document corpus exceeds model context lengths. Therefore the use of out-of-context retrieval, such as RAG prompting and agentic prompting techniques, is necessary for obtaining the answers. Here we observe that RAG prompting techniques, which select relevant documents for the question using vector embeddings, deliver poor F1 scores for all universe sizes—the performance only deteriorates with increasing universe size. Agentic prompting techniques like ReAct show immense promise by avoiding a steep downward trend. This suggests that **agentic workflows can be effective in dynamically retrieving documents at scale**.

7 Conclusion and Future Work

We introduce *PhantomWiki*: a benchmarking framework to evaluate reasoning and retrieval capabilities of language models. As we increase the question complexity and universe size, we observe that current state-of-the-art LLMs struggle in both reasoning and retrieval aspects. *PhantomWiki* is scalable and memorization-resistant, hence well-suited to evaluate future generations of language models.

Our work brings forth several research directions. Noting how we generate document corpora and questions, *PhantomWiki* is resistant to data contamination. We leave to future work to empirically test this claim, and develop theory to formally prove that our benchmark is memorization-resistant. In this work we focus on question-answering over text corpora. We hope to extend *PhantomWiki* for other knowledge bases and modalities such as vision and audio, enabling analogous evaluation suites for multimodal models.

References

AlphaProof & AlphaGeometry. AI achieves silver-medal standard solving International Mathematical Olympiad problems. https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/, 2024. Accessed: 2025-01-25.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.

Yair Feldman and Ran El-Yaniv. Multi-hop paragraph retrieval for open-domain question answering. *arXiv preprint arXiv:1906.06606*, 2019.

P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009. ISBN 9781139477161. URL `https://books.google.com/books?id=0h-4QcA1c1QC`.

Gemini Team, Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, et al. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Patrick Hohenecker and Thomas Lukasiewicz. Ontology reasoning with deep neural networks. *Journal of Artificial Intelligence Research*, 68:503–540, 2020.

John E Hopcroft, Rajeev Motwani, and Jeffrey D Ullman. Introduction to automata theory, languages, and computation. *Acm Sigact News*, 32(1):60–65, 2001.

Jennifer Hsia, Afreen Shaikh, Zhiruo Wang, and Graham Neubig. RAGGED: Towards informed design of retrieval augmented generation systems. *arXiv preprint arXiv:2403.09040*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jiajie Jin, Yutao Zhu, Xinyu Yang, Chenghao Zhang, and Zhicheng Dou. FlashRAG: A modular toolkit for efficient retrieval-augmented generation research. *arXiv preprint arXiv:2405.13576*, 2024.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

Barrett Martin Lattimer, Varun Gangal, Ryan McDonald, and Yi Yang. Sparse rewards can self-train dialogue agents. *arXiv preprint arXiv:2409.04617*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Qianren Mao, Yangyifei Luo, Jinlong Zhang, Hanwen Hao, Zhilong Cao, Xiaolong Wang, Xiao Guan, Zhenting Huang, Weifeng Jiang, Shuyu Guo, et al. XRAG: eXamining the core–benchmarking foundational components in advanced retrieval-augmented generation. *arXiv preprint arXiv:2412.15529*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. A discrete hard em approach for weakly supervised question answering. *arXiv preprint arXiv:1909.04849*, 2019.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive APIs. *arXiv preprint arXiv:2305.15334*, 2023.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*, 2020.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Vassilina Nikoulina, and Stéphane Clinchant. BERGEN: A benchmarking library for retrieval-augmented generation. *arXiv preprint arXiv:2407.01102*, 2024.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*, 2023.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*, 2022.

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. Assisting in writing Wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. *arXiv preprint arXiv:1908.06177*, 2019.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*, 2023.

Leon Sterling and Ehud Y Shapiro. *The art of Prolog: advanced programming techniques*. MIT press, 1994.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, et al. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*, 2024.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. *arXiv preprint arXiv:2012.13048*, 2020.

Yixuan Tang and Yi Yang. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen 2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. WebArena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. Large language models can learn rules. *arXiv preprint arXiv:2310.07064*, 2023.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. ToolQA: A dataset for LLM question answering with external tools. *Advances in Neural Information Processing Systems*, 36: 50117–50143, 2023.

A Background

A.1 Context-Free Grammars

Context-free grammar (CFG) is a type of formal grammar where the productions rules govern how to generate text from non-terminals and terminals. A context-free grammar is defined by $G = (V, \Sigma, R, S)$ where V and Σ denotes nonterminal and terminal respectively. R is a finite relation in $V \times (V \cup \Sigma)^*$ which specifies the production rules of the grammar. $S \in V$ is the start symbol. A production rule in R has the form

$$\alpha \to \beta \tag{1}$$

where $\alpha \in V$, $\beta \in (V \cup \Sigma)^*$. It is conventional to list all rules with the same left-hand side on the same line and separate the right-hand side with "|" like $\alpha \to \beta_1 \mid \beta_2$.

B Question template generation

B.1 Context-Free Grammar

We use the following CFG to generate question templates:

```
S -> Who is R? | What is A ? | How many RN_p does R_c have ?
R -> the RN of R_c | the person whose AN is AV
R_c -> R | N
A -> the AN of R
RN -> <relation>
RN_p -> <relation_plural>
AN -> <attribute_name>
AV -> <attribute_value>
N -> <name>
```

B.2 CFG-generated question templates

Our CFG produces the following 50 question templates at recursion limit $d = 20$. Note how the recursive production rule `R_c -> R | N` leads to chained productions.

```
1. Who is the <relation>_3 of the <relation>_5 of the <relation>_7
   of the <relation>_9 of the <relation>_11 of the <relation>_13
   of the <relation>_15 of the <relation>_17 of the person whose
   <attribute_name>_19 is <attribute_value>_19?
2. Who is the <relation>_3 of the <relation>_5 of the <relation>_7
   of the <relation>_9 of the <relation>_11 of the <relation>_13
   of the <relation>_15 of the <relation>_17 of <name>_18?
3. Who is the <relation>_3 of the <relation>_5 of the <relation>_7
   of the <relation>_9 of the <relation>_11 of the <relation>_13
   of the <relation>_15 of the person whose <attribute_name>_17
   is <attribute_value>_17?
4. Who is the <relation>_3 of the <relation>_5 of the <relation>_7
   of the <relation>_9 of the <relation>_11 of the <relation>_13
   of the <relation>_15 of <name>_16?
5. Who is the <relation>_3 of the <relation>_5 of the <relation>_7
   of the <relation>_9 of the <relation>_11 of the <relation>_13
   of the person whose <attribute_name>_15 is <attribute_value>
   _15?
6. Who is the <relation>_3 of the <relation>_5 of the <relation>_7
   of the <relation>_9 of the <relation>_11 of the <relation>_13
   of <name>_14?
7. Who is the <relation>_3 of the <relation>_5 of the <relation>_7
   of the <relation>_9 of the <relation>_11 of the person whose
   <attribute_name>_13 is <attribute_value>_13?
```

```
8. Who is the <relation>_3 of the <relation>_5 of the <relation>_7
    of the <relation>_9 of the <relation>_11 of <name>_12?
9. Who is the <relation>_3 of the <relation>_5 of the <relation>_7
    of the <relation>_9 of the person whose <attribute_name>_11
   is <attribute_value>_11?
10. Who is the <relation>_3 of the <relation>_5 of the <relation>
   _7 of the <relation>_9 of <name>_10?
11. Who is the <relation>_3 of the <relation>_5 of the <relation>
   _7 of the person whose <attribute_name>_9 is <attribute_value>
   _9?
12. Who is the <relation>_3 of the <relation>_5 of the <relation>
   _7 of <name>_8?
13. Who is the <relation>_3 of the <relation>_5 of the person
   whose <attribute_name>_7 is <attribute_value>_7?
14. Who is the <relation>_3 of the <relation>_5 of <name>_6?
15. Who is the <relation>_3 of the person whose <attribute_name>_5
    is <attribute_value>_5?
16. Who is the <relation>_3 of <name>_4?
17. Who is the person whose <attribute_name>_3 is <attribute_value
   >_3?
18. What is the <attribute_name>_3 of the <relation>_4 of the <
   relation>_6 of the <relation>_8 of the <relation>_10 of the <
   relation>_12 of the <relation>_14 of the <relation>_16 of the
   <relation>_18 of <name>_19?
19. What is the <attribute_name>_3 of the <relation>_4 of the <
   relation>_6 of the <relation>_8 of the <relation>_10 of the <
   relation>_12 of the <relation>_14 of the <relation>_16 of the
   person whose <attribute_name>_18 is <attribute_value>_18?
20. What is the <attribute_name>_3 of the <relation>_4 of the <
   relation>_6 of the <relation>_8 of the <relation>_10 of the <
   relation>_12 of the <relation>_14 of the <relation>_16 of <
   name>_17?
21. What is the <attribute_name>_3 of the <relation>_4 of the <
   relation>_6 of the <relation>_8 of the <relation>_10 of the <
   relation>_12 of the <relation>_14 of the person whose <
   attribute_name>_16 is <attribute_value>_16?
22. What is the <attribute_name>_3 of the <relation>_4 of the <
   relation>_6 of the <relation>_8 of the <relation>_10 of the <
   relation>_12 of the <relation>_14 of <name>_15?
23. What is the <attribute_name>_3 of the <relation>_4 of the <
   relation>_6 of the <relation>_8 of the <relation>_10 of the <
   relation>_12 of the person whose <attribute_name>_14 is <
   attribute_value>_14?
24. What is the <attribute_name>_3 of the <relation>_4 of the <
   relation>_6 of the <relation>_8 of the <relation>_10 of the <
   relation>_12 of <name>_13?
25. What is the <attribute_name>_3 of the <relation>_4 of the <
   relation>_6 of the <relation>_8 of the <relation>_10 of the
   person whose <attribute_name>_12 is <attribute_value>_12?
26. What is the <attribute_name>_3 of the <relation>_4 of the <
   relation>_6 of the <relation>_8 of the <relation>_10 of <name>
   _11?
27. What is the <attribute_name>_3 of the <relation>_4 of the <
   relation>_6 of the <relation>_8 of the person whose <
   attribute_name>_10 is <attribute_value>_10?
28. What is the <attribute_name>_3 of the <relation>_4 of the <
   relation>_6 of the <relation>_8 of <name>_9?
```

29. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of the person whose <attribute_name>_8 is <attribute_value>_8?
30. What is the <attribute_name>_3 of the <relation>_4 of the <relation>_6 of <name>_7?
31. What is the <attribute_name>_3 of the <relation>_4 of the person whose <attribute_name>_6 is <attribute_value>_6?
32. What is the <attribute_name>_3 of the <relation>_4 of <name>_5?
33. What is the <attribute_name>_3 of the person whose <attribute_name>_4 is <attribute_value>_4?
34. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the <relation>_16 of the <relation>_18 of <name>_19 have?
35. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the <relation>_16 of the person whose <attribute_name>_18 is <attribute_value>_18 have?
36. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the <relation>_16 of <name>_17 have?
37. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of the person whose <attribute_name>_16 is <attribute_value>_16 have?
38. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the <relation>_14 of <name>_15 have?
39. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of the person whose <attribute_name>_14 is <attribute_value>_14 have?
40. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the <relation>_12 of <name>_13 have?
41. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of the person whose <attribute_name>_12 is <attribute_value>_12 have?
42. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the <relation>_10 of <name>_11 have?
43. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of the person whose <attribute_name>_10 is <attribute_value>_10 have?
44. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the <relation>_8 of <name>_9 have?
45. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of the person whose <attribute_name>_8 is <attribute_value>_8 have?
46. How many <relation_plural>_2 does the <relation>_4 of the <relation>_6 of <name>_7 have?
47. How many <relation_plural>_2 does the <relation>_4 of the person whose <attribute_name>_6 is <attribute_value>_6 have?
48. How many <relation_plural>_2 does the <relation>_4 of <name>_5 have?
49. How many <relation_plural>_2 does the person whose <attribute_name>_4 is <attribute_value>_4 have?

50. How many <relation_plural>_2 does <name>_3 have?

B.3 Question-Answer Characteristics

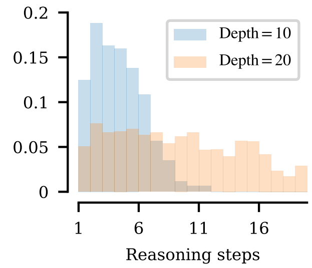

Figure 5: Histogram of question difficulties (measured by reasoning steps) for universe size $n = 50$ at two CFG recursion limits $d \in \{10, 20\}$. We average the frequencies across 3 dataset generation seeds.

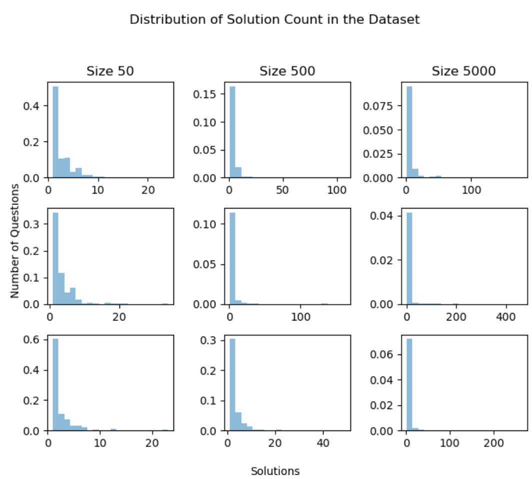

Figure 6: Distribution of number of answers across sizes $n \in \{50, 500, 5000\}$, seeds $\{1, 2, 3\}$, and CFG depth 20.

C Baseline Details

C.1 LLM Sampling Hyperparameters

	Temperature	Top-k	Top-p	Repetition Penalty	Sampling Seed	Max number of output Tokens
Values	0	50	0.7	1.0	0	4096

Table 3: Default Hyperparameters values for LLM Sampling

We used the above default hyperparameters values for all models, but DeepSeek-R1-32B, where we used temperature $= 0.6$ and top-p $= 0.95$.

C.2 ZeroShot-Simple

We use the following prompt for all models, where `evidence` is the concatenation of all documents in the *PhantomWiki* instance.

```
You are given the following evidence:
(BEGIN EVIDENCE)
{{evidence}}
(END EVIDENCE)

You will be provided a question. Your task is to provide an answer
     according to these instructions:
- The output must be one of the following: a name (if there is
    only one correct answer); or a list of names separated by '{
    constants.answer_sep}' (if there are multiple correct answers)
    .
- DO NOT include any additional information in your answer.

Question: {{question}}
Answer:
```

For DeepSeek-R1-32B, we additionally parse the output to separate the model's reasoning process from its final answer using the `</think>` tag.

C.3 ZEROSHOT-RAG

The prompt is exactly the same as ZEROSHOT, except we replace `evidence` with 4 documents retrieved using the UAE-LARGE-V1. We pre-compute an index the document corpus using an FAISS vector store of UAE-LARGE-V1 embeddings. Upon generation, we search for similar documents for `question` according to maximum inner product search on document and question embeddings.

C.4 CHAIN-OF-THOUGHT-SIMPLE

We use the following prompt for all models, where `evidence` is replaced with a list of all documents. We use a regular expression to parse the output.

```
You are given the following evidence:
(BEGIN EVIDENCE)
{{evidence}}
(END EVIDENCE)

You will be provided a question. Your response must end in the
    following sentence: The answer is <answer>.
Here, <answer> must be one of the following:
- a name (if there is only one correct answer); or
- a list of names separated by '{constants.answer_sep}' (if there
    are multiple correct answers).

Here are some examples:
(START OF EXAMPLES)
Example 1:
Question: Who is the brother of Dino Beltran?
Answer: Based on the evidence, the brother of Dino Beltran is
    Orlando Beltran. The answer is Orlando Beltran.

Example 2:
Question: Who is the sibling of Barabara Beltran?
Answer: Based on the evidence, the siblings of Barabara Beltran
    are Aida Wang, Vicki Hackworth. The answer is Aida Wang{
    constants.answer_sep}Vicki Hackworth.

Example 3:
Question: Who is the child of the sibling of Stacia Toombs?
```

```
Answer: First I need to find the sibling of Stacia Toombs. Based
    on the evidence, the sibling of Stacia Toombs is Shelli
    Beltran. Now I need to find the child of Shelli Beltran. Based
     on the evidence, the children of Shelli Beltran are Aida Wang
    , Barabara Beltran, Vicki Hackworth. The answer is Aida Wang{
    constants.answer_sep}Barabara Beltran{constants.answer_sep}
    Vicki Hackworth.

Example 4:
Question: Who is the uncle of William Smock?
Answer: An uncle is the brother of a parent. Based on the evidence
    , the parents of William Smock are Dominique Smock, Gene Smock
    . To find the uncle of William Smock, I need to find the
    brother of Dominique Smock and Gene Smock. Based on the
    evidence, Dominique Smock has no brother, and the brother of
    Gene Smock is Eli Smock. So the uncle of William Smock is Eli
    Smock. The answer is Eli Smock.

Example 5:
Question: What is the occupation of the sister of the grandmother
    of Virgil Hackworth?
Answer: A grandmother is the mother of a parent. Based on the
    evidence, the parents of Virgil Hackworth are Ricardo
    Hackworth, Vicki Hackworth. To find the grandmother of Virgil
    Hackworth, I need to find the mother of Ricardo Hackworth and
    Vicki Hackworth. Based on the evidence, Ricardo Hackworth has
    no mother, and the mother of Vicki Hackworth is Shelli Beltran
    . Now I need to find the sister of Shelli Beltran. Based on
    the evidence, the sister of Shelli Beltran is Stacia Toombs.
    Based on the evidence, the occupation of Stacia Toombs is
    actuary. The answer is actuary.

Example 6:
Question: Who is the brother of the person whose occupation is
    associate professor?
Answer: I need to search for people whose occupation is associate
    professor. Based on the evidence, the person whose occupation
    is associate professor is Dino Beltran. And the brother of
    Dino Beltran is Orlando Beltran. The answer is Orlando Beltran
    .

Example 7:
Question: What is the date of birth of the person whose hobby is
    meteorology?
Answer: I need to search for people whose hobby is meteorology.
    Based on the evidence, the people whose hobby is meteorology
    are Alison Smock, Barabara Beltran. The date of birth of
    Alison Smock is 0929-10-28, and the date of birth of Barabara
    Beltran is 0989-06-11. The answer is 0929-10-28{constants.
    answer_sep}0989-06-11.

Example 8:
Question: Who is the cousin of the person whose occupation is
    broadcast engineer?
Answer: I need to search for people whose occupation is broadcast
    engineer. Based on the evidence, the person whose occupation
    is broadcast engineer is Barabara Beltran. A cousin is the
    child of the sibling of the parent. Based on the evidence, the
     parents of Barabara Beltran are Dino Beltran, Shelli Beltran.
```

```
    The sibling of Dino Beltran is Orlando Beltran, and the
   sibling of Shelli Beltran is Stacia Toombs. Based on the
   evidence, Orlando Beltran has no child, and the child of
   Stacia Toombs is Leslee Toombs. So the cousin of Barabara
   Beltran is Leslee Toombs. The answer is Leslee Toombs.

Example 9:
Question: Who is the great-granddaughter of the person whose hobby
    is biology?
Answer: I need to search for people whose hobby is biology. Based
   on the evidence, the person whose hobby is biology is Alvaro
   Smock. To find the great-granddaughter of Alvaro Smock, I need
    to find the daughter of the child of the child of Alvaro
   Smock. Based on the evidence, the children of Alvaro Smock are
    Eli Smock, Gene Smock. Eli Smock has no child, and the child
   of Gene Smock is Williams Smock. The daughters of Williams
   Smock are Shelli Beltran, Stacia Toombs. So the great-
   granddaughters of Alvaro Smock, whose hobby is biology, are
   Shelli Beltran, Stacia Toombs. The answer is Shelli Beltran{
   constants.answer_sep}Stacia Toombs.
(END OF EXAMPLES)

Question: {{question}}
Answer:
```

C.5 CHAIN-OF-THOUGHT-RAG

The prompt is exactly the same as COT, except we replace `evidence` with 4 documents retrieved using the UAE-LARGE-V1. We use the same retriever setup as described in Appendix C.3.

C.6 REACT

We modify the ReAct agent implementation of Shinn et al. (2024), which can be found at https://github.com/noahshinn/reflexion, to support our keyword-based search tools `RetrieveArticle` and `Search`. As seen in the following examples, the `RetrieveArticle` tool fetches the documents by title, and the `Search` tool fetches all document titles whose bodies contain the argument. We instruct the LLM to output predict the answer using a third tool `Finish`.

```
Solve a question answering task with interleaving Thought, Action,
    Observation steps.
Thought can reason about the current situation, and Action can be
   3 types:
(1) RetrieveArticle[{{{{entity}}}}]. This action retrieves the
   article about {{{{entity}}}} if it exists. If the article does
    not exist, the action will say so.
(2) Search[{{{{attribute}}}}]. This action searches the database
   for {{{{attribute}}}} and retrieves all articles that contain
   {{{{attribute}}}}. If no article contains {{{{attribute}}}},
   the action will say so.
(3) Finish[{{{{answer}}}}]. This action answers the question with
   {{{{answer}}}}.
If you cannot find the answer, output the empty answer like:
   Finish[].
If there are multiple answers A,B,C, answer with a list like:
   Finish[A{constants.answer_sep}B{constants.answer_sep}C].

You may take as many steps as necessary.
Here are some examples:
(START OF EXAMPLES)
```

```
Example 1:
Question: Who is the brother of Dino Beltran?
Thought 1: I need to retrieve article about Dino Beltran and find
   who his brother is.
Action 1: RetrieveArticle[Dino Beltran].
Observation 1: # Dino Beltran ## Family The parents of Dino
   Beltran are Brian Beltran, Daisy Beltran.  Dino Beltran's
   sibling is Orlando Beltran.  The brother of Dino Beltran is
   Orlando Beltran.  The mother of Dino Beltran is Daisy Beltran.
     The father of Dino Beltran is Brian Beltran.  The children
   of Dino Beltran are Aida Wang, Barabara Beltran, Vicki
   Hackworth.  The daughters of Dino Beltran are Aida Wang,
   Barabara Beltran, Vicki Hackworth.  The wife of Dino Beltran
   is Shelli Beltran.  ## Friends The friend of Dino Beltran is
   Alvaro Smock.  ## Attributes The date of birth of Dino Beltran
    is 0958-08-09.  The occupation of Dino Beltran is associate
   professor.  The hobby of Dino Beltran is shogi.
Thought 2: The brother of Dino Beltran is Orlando Beltran, so the
   answer is Orlando Beltran.
Action 2: Finish[Orlando Beltran].

Example 2:
Question: Who is the mother of Ivana Smith?
Thought 1: I need to retrieve article about Ivana Smith and find
   who her mother is.
Action 1: RetrieveArticle[Ivana Smith].
Observation 1: No article exists for the requested entity. Please
   try retrieving article for another entity.
Thought 2: Article about Ivana Smith does not exist, so I cannot
   find the mother of Ivana Smith. I will output empty answer.
Action 2: Finish[].

Example 3:
Question: Who is the sibling of Barabara Beltran?
Thought 1: I need to retrieve article about Barabara Beltran and
   find who her sibling is.
Action 1: RetrieveArticle[Barabara Beltran].
Observation 1: # Barabara Beltran ## Family The parents of
   Barabara Beltran are Dino Beltran, Shelli Beltran.  Barabara
   Beltran's siblings are Aida Wang, Vicki Hackworth.  The
   sisters of Barabara Beltran are Aida Wang, Vicki Hackworth.
   The mother of Barabara Beltran is Shelli Beltran.  The father
   of Barabara Beltran is Dino Beltran.  ## Friends ## Attributes
    The date of birth of Barabara Beltran is 0989-06-11.  The
   occupation of Barabara Beltran is broadcast engineer.  The
   hobby of Barabara Beltran is meteorology.
Thought 2: The siblings of Barabara Beltran are Aida Wang, Vicki
   Hackworth. So the answer is Aida Wang, Vicki Hackworth.
Action 2: Finish[Aida Wang{constants.answer_sep}Vicki Hackworth].

Example 4:
Question: Who is the child of the sibling of Stacia Toombs?
Thought 1: First, I need to retrieve article about Stacia Toombs
   and find who her sibling is.
Action 1: RetrieveArticle[Stacia Toombs].
Observation 1: # Stacia Toombs ## Family The parents of Stacia
   Toombs are Alison Smock, Williams Smock.  Stacia Toombs's
   sibling is Shelli Beltran.  The sister of Stacia Toombs is
   Shelli Beltran.  The mother of Stacia Toombs is Alison Smock.
```

```
 The father of Stacia Toombs is Williams Smock.  The child of
   Stacia Toombs is Leslee Toombs.  The daughter of Stacia Toombs
    is Leslee Toombs.  The husband of Stacia Toombs is Wilbert
   Toombs.  ## Friends The friends of Stacia Toombs are Brian
   Beltran, Isiah Lutz, Leeann Hackworth, Lesley Lutz, Ryan Wang.
     ## Attributes The date of birth of Stacia Toombs is
   0959-03-22.  The occupation of Stacia Toombs is actuary.  The
   hobby of Stacia Toombs is finance.
Thought 2: The sibling of Stacia Toombs is Shelli Beltran. Now I
   need to retrieve article about Shelli Beltran and find who her
    child is.
Action 2: RetrieveArticle[Shelli Beltran].
Observation 2: # Shelli Beltran ## Family The parents of Shelli
   Beltran are Alison Smock, Williams Smock.  Shelli Beltran's
   sibling is Stacia Toombs.  The sister of Shelli Beltran is
   Stacia Toombs.  The mother of Shelli Beltran is Alison Smock.
    The father of Shelli Beltran is Williams Smock.  The children
    of Shelli Beltran are Aida Wang, Barabara Beltran, Vicki
   Hackworth.  The daughters of Shelli Beltran are Aida Wang,
   Barabara Beltran, Vicki Hackworth.  The husband of Shelli
   Beltran is Dino Beltran.  ## Friends The friends of Shelli
   Beltran are Brian Beltran, Eli Smock, Isiah Lutz, Leslee
   Toombs, Lesley Lutz, Ryan Wang.  ## Attributes The date of
   birth of Shelli Beltran is 0958-03-08.  The occupation of
   Shelli Beltran is occupational therapist.  The hobby of Shelli
    Beltran is sociology.
Thought 3: The child of Shelli Beltran is Aida Wang, Barabara
   Beltran, Vicki Hackworth. So the answer is Aida Wang, Barabara
    Beltran, Vicki Hackworth.
Action 3: Finish[Aida Wang{constants.answer_sep}Barabara Beltran{
   constants.answer_sep}Vicki Hackworth].

Example 5:
Question: Who is the uncle of William Smock?
Thought 1: First, I need to retrieve article about William Smock.
   To find the uncle of William Smock, I need to find the brother
    of his parent.
Action 1: RetrieveArticle[William Smock].
Observation 1: # Williams Smock ## Family The parents of Williams
   Smock are Dominique Smock, Gene Smock.  The mother of Williams
    Smock is Dominique Smock.  The father of Williams Smock is
   Gene Smock.  The children of Williams Smock are Shelli Beltran
   , Stacia Toombs.  The daughters of Williams Smock are Shelli
   Beltran, Stacia Toombs.  The wife of Williams Smock is Alison
   Smock.  ## Friends The friend of Williams Smock is Lannie
   Smock.  ## Attributes The date of birth of Williams Smock is
   0926-04-04.  The occupation of Williams Smock is clinical
   biochemist.  The hobby of Williams Smock is social studies.
Thought 2: The parents of William Smock are Dominique Smock, Gene
   Smock. To find the uncle of William Smock, I need to find the
   brother of Dominique Smock and Gene Smock.
Action 2: RetrieveArticle[Dominique Smock].
Observation 2: # Dominique Smock ## Family The child of Dominique
   Smock is Williams Smock.  The son of Dominique Smock is
   Williams Smock.  The husband of Dominique Smock is Gene Smock.
     ## Friends The friends of Dominique Smock are Gene Smock,
   Isiah Lutz, Orlando Beltran, Vicki Hackworth.  ## Attributes
   The date of birth of Dominique Smock is 0897-09-08.  The
```

```
    occupation of Dominique Smock is sports therapist.  The hobby
    of Dominique Smock is dominoes.
Thought 3: Dominique Smock has no brother. Now I need to retrieve
    article about Gene Smock and find his brother.
Action 3: RetrieveArticle[Gene Smock].
Observation 3: # Gene Smock ## Family The parents of Gene Smock
    are Alvaro Smock, Lannie Smock.  Gene Smock's sibling is Eli
    Smock.  The brother of Gene Smock is Eli Smock.  The mother of
     Gene Smock is Lannie Smock.  The father of Gene Smock is
    Alvaro Smock.  The child of Gene Smock is Williams Smock.  The
     son of Gene Smock is Williams Smock.  The wife of Gene Smock
    is Dominique Smock.  ## Friends The friends of Gene Smock are
    Leeann Hackworth, Leisa Lutz, Ricardo Hackworth, Alvaro Smock,
     Dominique Smock.  ## Attributes The date of birth of Gene
    Smock is 0898-08-16.  The occupation of Gene Smock is
    immunologist.  The hobby of Gene Smock is architecture.
Thought 4: The brother of Gene Smock is Eli Smock, and Dominique
    Smock has no brother. Since Gene Smock and Dominique Smock are
     the parents of William Smock, the uncle of William Smock is
    Eli Smock. So the answer is Eli Smock.
Action 4: Finish[Eli Smock].

Example 6:
Question: What is the occupation of the sister of the grandmother
    of Virgil Hackworth?
Thought 1: First, I need to retrieve article about Virgil
    Hackworth. To find the grandmother of Virgil Hackworth, I need
     to find the mother of his parent.
Action 1: RetrieveArticle[Virgil Hackworth].
Observation 1: # Virgil Hackworth ## Family The parents of Virgil
    Hackworth are Ricardo Hackworth, Vicki Hackworth.  Virgil
    Hackworth's siblings are Leeann Hackworth, Leisa Lutz.  The
    sisters of Virgil Hackworth are Leeann Hackworth, Leisa Lutz.
     The mother of Virgil Hackworth is Vicki Hackworth.  The
    father of Virgil Hackworth is Ricardo Hackworth.  ## Friends
    The friends of Virgil Hackworth are Alison Smock, Eli Smock,
    Orlando Beltran, Ryan Wang.  ## Attributes The date of birth
    of Virgil Hackworth is 1009-06-23.  The occupation of Virgil
    Hackworth is sound technician.  The hobby of Virgil Hackworth
    is wikipedia editing.
Thought 2: The parents of Virgil Hackworth are Ricardo Hackworth,
    Vicki Hackworth. To find the grandmother of Virgil Hackworth,
    I need to find the mother of Ricardo Hackworth and Vicki
    Hackworth.
Action 2: RetrieveArticle[Ricardo Hackworth].
Observation 2: # Ricardo Hackworth ## Family The children of
    Ricardo Hackworth are Leeann Hackworth, Leisa Lutz, Virgil
    Hackworth.  The son of Ricardo Hackworth is Virgil Hackworth.
     The daughters of Ricardo Hackworth are Leeann Hackworth,
    Leisa Lutz.  The wife of Ricardo Hackworth is Vicki Hackworth.
      ## Friends The friends of Ricardo Hackworth are Gene Smock,
    Isiah Lutz, Johnetta Wang.  ## Attributes The date of birth of
     Ricardo Hackworth is 0983-02-24.  The occupation of Ricardo
    Hackworth is clinical research associate.  The hobby of
    Ricardo Hackworth is dairy farming.
Thought 3: Ricardo Hackworth has no mother. So I need to retrieve
    article about Vicki Hackworth and find her mother.
Action 3: RetrieveArticle[Vicki Hackworth].
```

```
Observation 3: # Vicki Hackworth ## Family The parents of Vicki
    Hackworth are Dino Beltran, Shelli Beltran.  Vicki Hackworth's
     siblings are Aida Wang, Barabara Beltran.  The sisters of
    Vicki Hackworth are Aida Wang, Barabara Beltran.  The mother
    of Vicki Hackworth is Shelli Beltran.  The father of Vicki
    Hackworth is Dino Beltran.  The children of Vicki Hackworth
    are Leeann Hackworth, Leisa Lutz, Virgil Hackworth.  The son
    of Vicki Hackworth is Virgil Hackworth.  The daughters of
    Vicki Hackworth are Leeann Hackworth, Leisa Lutz.  The husband
     of Vicki Hackworth is Ricardo Hackworth.  ## Friends The
    friends of Vicki Hackworth are Brian Beltran, Dominique Smock,
     Eli Smock.  ## Attributes The date of birth of Vicki
    Hackworth is 0985-05-30.  The occupation of Vicki Hackworth is
     police officer.  The hobby of Vicki Hackworth is meditation.
Thought 4: The mother of Vicki Hackworth is Shelli Beltran. So the
     only grandmother of Virgil Hackworth is Shelli Beltran. Now I
     need to retrieve article about Shelli Beltran and find her
    sister.
Action 4: RetrieveArticle[Shelli Beltran].
Observation 4: # Shelli Beltran ## Family The parents of Shelli
    Beltran are Alison Smock, Williams Smock.  Shelli Beltran's
    sibling is Stacia Toombs.  The sister of Shelli Beltran is
    Stacia Toombs.  The mother of Shelli Beltran is Alison Smock.
     The father of Shelli Beltran is Williams Smock.  The children
     of Shelli Beltran are Aida Wang, Barabara Beltran, Vicki
    Hackworth.  The daughters of Shelli Beltran are Aida Wang,
    Barabara Beltran, Vicki Hackworth.  The husband of Shelli
    Beltran is Dino Beltran.  ## Friends The friends of Shelli
    Beltran are Brian Beltran, Eli Smock, Isiah Lutz, Leslee
    Toombs, Lesley Lutz, Ryan Wang.  ## Attributes The date of
    birth of Shelli Beltran is 0958-03-08.  The occupation of
    Shelli Beltran is occupational therapist.  The hobby of Shelli
     Beltran is sociology.
Thought 5: The sister of Shelli Beltran is Stacia Toombs. Now I
    need to retrieve article about Stacia Toombs and find her
    occupation.
Action 5: RetrieveArticle[Stacia Toombs].
Observation 5: # Stacia Toombs ## Family The parents of Stacia
    Toombs are Alison Smock, Williams Smock.  Stacia Toombs's
    sibling is Shelli Beltran.  The sister of Stacia Toombs is
    Shelli Beltran.  The mother of Stacia Toombs is Alison Smock.
     The father of Stacia Toombs is Williams Smock.  The child of
    Stacia Toombs is Leslee Toombs.  The daughter of Stacia Toombs
     is Leslee Toombs.  The husband of Stacia Toombs is Wilbert
    Toombs.  ## Friends The friends of Stacia Toombs are Brian
    Beltran, Isiah Lutz, Leeann Hackworth, Lesley Lutz, Ryan Wang.
      ## Attributes The date of birth of Stacia Toombs is
    0959-03-22.  The occupation of Stacia Toombs is actuary.  The
    hobby of Stacia Toombs is finance.
Thought 6: The occupation of Stacia Toombs is actuary. So the
    answer is actuary.
Action 6: Finish[actuary].

Example 7:
Question: Who is the brother of the person whose occupation is
    associate professor?
Thought 1: First, I need to search for people whose occupation is
    associate professor.
Action 1: Search[associate professor].
```

```
Observation 1: (1) Dino Beltran
Thought 2: Person whose hobby is associate professor is Dino
   Beltran. Now I need to retrieve article about Dino Beltran and
    find his brother.
Action 2: RetrieveArticle[Dino Beltran].
Observation 2: # Dino Beltran ## Family The parents of Dino
   Beltran are Brian Beltran, Daisy Beltran.  Dino Beltran’s
   sibling is Orlando Beltran.  The brother of Dino Beltran is
   Orlando Beltran.  The mother of Dino Beltran is Daisy Beltran.
     The father of Dino Beltran is Brian Beltran.  The children
   of Dino Beltran are Aida Wang, Barabara Beltran, Vicki
   Hackworth.  The daughters of Dino Beltran are Aida Wang,
   Barabara Beltran, Vicki Hackworth.  The wife of Dino Beltran
   is Shelli Beltran.  ## Friends The friend of Dino Beltran is
   Alvaro Smock.  ## Attributes The date of birth of Dino Beltran
    is 0958-08-09.  The occupation of Dino Beltran is associate
   professor.  The hobby of Dino Beltran is shogi.
Thought 3: The brother of Dino Beltran is Orlando Beltran, so the
   answer is Orlando Beltran.
Action 3: Finish[Orlando Beltran].

Example 8:
Question: What is the date of birth of the person whose hobby is
   meteorology?
Thought 1: First, I need to search for people whose hobby is
   meteorology.
Action 1: Search[meteorology].
Observation 1: (1) Alison Smock (2) Barabara Beltran
Thought 2: People whose hobby is meteorology are Alison Smock,
   Barabara Beltran. Now I need to retrieve article about Alison
   Smock and Barabara Beltran and find their date of birth.
Action 2: RetrieveArticle[Alison Smock].
Observation 2: # Alison Smock ## Family The children of Alison
   Smock are Shelli Beltran, Stacia Toombs.  The daughters of
   Alison Smock are Shelli Beltran, Stacia Toombs.  The husband
   of Alison Smock is Williams Smock.  ## Friends The friends of
   Alison Smock are Eli Smock, Lannie Smock, Virgil Hackworth.
   ## Attributes The date of birth of Alison Smock is 0929-10-28.
     The occupation of Alison Smock is health promotion
   specialist.  The hobby of Alison Smock is meteorology.
Thought 3: The date of birth of Alison Smock is 0929-10-28. Now I
   need to retrieve article about Barabara Beltran and find her
   date of birth.
Action 3: RetrieveArticle[Barabara Beltran].
Observation 3: # Barabara Beltran ## Family The parents of
   Barabara Beltran are Dino Beltran, Shelli Beltran.  Barabara
   Beltran’s siblings are Aida Wang, Vicki Hackworth.  The
   sisters of Barabara Beltran are Aida Wang, Vicki Hackworth.
   The mother of Barabara Beltran is Shelli Beltran.  The father
   of Barabara Beltran is Dino Beltran.  ## Friends ## Attributes
    The date of birth of Barabara Beltran is 0989-06-11.  The
   occupation of Barabara Beltran is broadcast engineer.  The
   hobby of Barabara Beltran is meteorology.
Thought 4: The date of birth of Barabara Beltran is 0989-06-11. So
    the answer is 0929-10-28, 0989-06-11.
Action 4: Finish[0929-10-28{constants.answer_sep}0989-06-11].

Example 9:
```

```
Question: Who is the cousin of the person whose occupation is
    broadcast engineer?
Thought 1: First, I need to search for people whose occupation is
    broadcast engineer.
Action 1: Search[broadcast engineer].
Observation 1: # Barabara Beltran ## Family The parents of
    Barabara Beltran are Dino Beltran, Shelli Beltran.  Barabara
    Beltran's siblings are Aida Wang, Vicki Hackworth.  The
    sisters of Barabara Beltran are Aida Wang, Vicki Hackworth.
    The mother of Barabara Beltran is Shelli Beltran.  The father
    of Barabara Beltran is Dino Beltran.  ## Friends ## Attributes
     The date of birth of Barabara Beltran is 0989-06-11.  The
    occupation of Barabara Beltran is broadcast engineer.  The
    hobby of Barabara Beltran is meteorology.
Thought 2: Person whose hobby is broadcast engineer is Barabara
    Beltran. The cousin of Barabara Beltran is the child of
    sibling of her parent. Barabara Beltran's parents are Dino
    Beltran and Shelly Beltran. Now I need to find the sibling of
    Dino Beltran and Shelly Beltran.
Action 2: RetrieveArticle[Dino Beltran].
Observation 2: # Dino Beltran ## Family The parents of Dino
    Beltran are Brian Beltran, Daisy Beltran.  Dino Beltran's
    sibling is Orlando Beltran.  The brother of Dino Beltran is
    Orlando Beltran.  The mother of Dino Beltran is Daisy Beltran.
      The father of Dino Beltran is Brian Beltran.  The children
    of Dino Beltran are Aida Wang, Barabara Beltran, Vicki
    Hackworth.  The daughters of Dino Beltran are Aida Wang,
    Barabara Beltran, Vicki Hackworth.  The wife of Dino Beltran
    is Shelli Beltran.  ## Friends The friend of Dino Beltran is
    Alvaro Smock.  ## Attributes The date of birth of Dino Beltran
     is 0958-08-09.  The occupation of Dino Beltran is associate
    professor.  The hobby of Dino Beltran is shogi.
Thought 3: The sibling of Dino Beltran is Orlando Beltran. Now I
    need to retrieve article about Orlando Beltran and find his
    child.
Action 3: RetrieveArticle[Orlando Beltran].
Observation 3: # Orlando Beltran ## Family The parents of Orlando
    Beltran are Brian Beltran, Daisy Beltran.  Orlando Beltran's
    sibling is Dino Beltran.  The brother of Orlando Beltran is
    Dino Beltran.  The mother of Orlando Beltran is Daisy Beltran.
      The father of Orlando Beltran is Brian Beltran.  ## Friends
    The friends of Orlando Beltran are Virgil Hackworth, Dominique
     Smock.  ## Attributes The date of birth of Orlando Beltran is
     0953-10-23.  The occupation of Orlando Beltran is petroleum
    engineer.  The hobby of Orlando Beltran is learning.
Thought 4: Orlando Beltran has no child, so no cousin from the
    side of Dino Beltran. Now I need to retrieve article about
    Shelli Beltran and find her sibling.
Action 4: RetrieveArticle[Shelli Beltran].
Observation 4: # Shelli Beltran ## Family The parents of Shelli
    Beltran are Alison Smock, Williams Smock.  Shelli Beltran's
    sibling is Stacia Toombs.  The sister of Shelli Beltran is
    Stacia Toombs.  The mother of Shelli Beltran is Alison Smock.
     The father of Shelli Beltran is Williams Smock.  The children
     of Shelli Beltran are Aida Wang, Barabara Beltran, Vicki
    Hackworth.  The daughters of Shelli Beltran are Aida Wang,
    Barabara Beltran, Vicki Hackworth.  The husband of Shelli
    Beltran is Dino Beltran.  ## Friends The friends of Shelli
    Beltran are Brian Beltran, Eli Smock, Isiah Lutz, Leslee
```

```
    Toombs, Lesley Lutz, Ryan Wang.  ## Attributes The date of
    birth of Shelli Beltran is 0958-03-08.  The occupation of
    Shelli Beltran is occupational therapist.  The hobby of Shelli
     Beltran is sociology.
Thought 5: The sibling of Shelli Beltran is Stacia Toombs. Now I
    need to retrieve article about Stacia Toombs and find her
    child, to get the cousin of Barabara Beltran.
Action 5: RetrieveArticle[Stacia Toombs].
Observation 5: # Stacia Toombs ## Family The parents of Stacia
    Toombs are Alison Smock, Williams Smock.  Stacia Toombs's
    sibling is Shelli Beltran.  The sister of Stacia Toombs is
    Shelli Beltran.  The mother of Stacia Toombs is Alison Smock.
     The father of Stacia Toombs is Williams Smock.  The child of
    Stacia Toombs is Leslee Toombs.  The daughter of Stacia Toombs
     is Leslee Toombs.  The husband of Stacia Toombs is Wilbert
    Toombs.  ## Friends The friends of Stacia Toombs are Brian
    Beltran, Isiah Lutz, Leeann Hackworth, Lesley Lutz, Ryan Wang.
      ## Attributes The date of birth of Stacia Toombs is
    0959-03-22.  The occupation of Stacia Toombs is actuary.  The
    hobby of Stacia Toombs is finance.
Thought 6: The child of Stacia Toombs is Leslee Toombs. So the
    cousin of Barabara Beltran is Leslee Toombs. The answer is
    Leslee Toombs.
Action 6: Finish[Leslee Toombs].

Example 10:
Question: Who is the great-granddaughter of the person whose hobby
     is biology?
Thought 1: First, I need to search for people whose hobby is
    biology.
Action 1: Search[biology].
Observation 1: (1) Alvaro Smock
Thought 2: Person whose hobby is biology is Alvaro Smock. Now I
    need to retrieve article about Alvaro Smock and find his great
    -granddaughter. A great-granddaughter is the daughter of the
    child of the child of the person.
Action 2: RetrieveArticle[Alvaro Smock].
Observation 2: # Alvaro Smock ## Family The children of Alvaro
    Smock are Eli Smock, Gene Smock.  The sons of Alvaro Smock are
     Eli Smock, Gene Smock.  The wife of Alvaro Smock is Lannie
    Smock.  ## Friends The friends of Alvaro Smock are Dino
    Beltran, Gene Smock, Aida Wang.  ## Attributes The date of
    birth of Alvaro Smock is 0867-07-12.  The occupation of Alvaro
     Smock is osteopath.  The hobby of Alvaro Smock is biology.
Thought 3: The children of Alvaro Smock are Eli Smock, Gene Smock.
     Now I need to retrieve article about Eli Smock and Gene Smock
     and find their child.
Action 3: RetrieveArticle[Eli Smock].
Observation 3: # Eli Smock ## Family The parents of Eli Smock are
    Alvaro Smock, Lannie Smock.  Eli Smock's sibling is Gene Smock
    .  The brother of Eli Smock is Gene Smock.  The mother of Eli
    Smock is Lannie Smock.  The father of Eli Smock is Alvaro
    Smock.  ## Friends The friends of Eli Smock are Leisa Lutz,
    Shelli Beltran, Vicki Hackworth, Virgil Hackworth, Alison
    Smock, Brian Beltran.  ## Attributes The date of birth of Eli
    Smock is 0901-01-18.  The occupation of Eli Smock is retail
    manager.  The hobby of Eli Smock is tether car.
Thought 4: Eli Smock has no child. Now I need to retrieve article
    about Gene Smock and find his child.
```

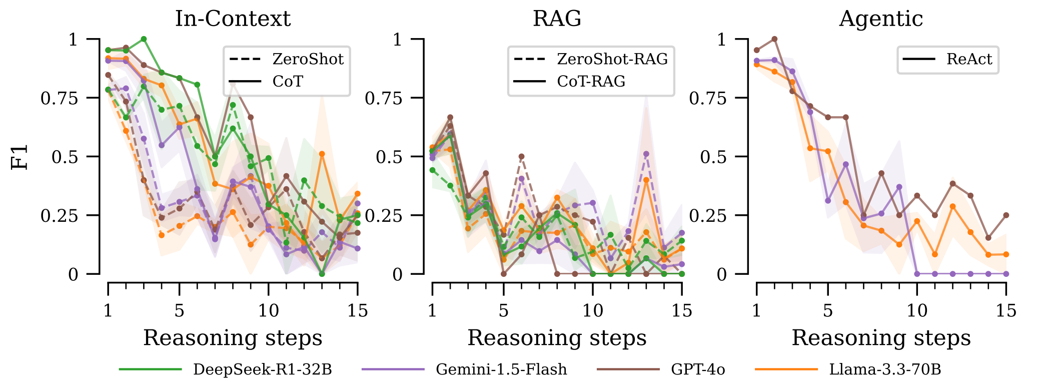

Figure 7: **F1 scores against *reasoning steps*, for questions with only 1 solution.**

```
Action 4: RetrieveArticle[Gene Smock].
Observation 4: # Gene Smock ## Family The parents of Gene Smock
    are Alvaro Smock, Lannie Smock.  Gene Smock's sibling is Eli
    Smock.  The brother of Gene Smock is Eli Smock.  The mother of
     Gene Smock is Lannie Smock.  The father of Gene Smock is
    Alvaro Smock.  The child of Gene Smock is Williams Smock.  The
     son of Gene Smock is Williams Smock.  The wife of Gene Smock
    is Dominique Smock.  ## Friends The friends of Gene Smock are
    Leeann Hackworth, Leisa Lutz, Ricardo Hackworth, Alvaro Smock,
     Dominique Smock.  ## Attributes The date of birth of Gene
    Smock is 0898-08-16.  The occupation of Gene Smock is
    immunologist.  The hobby of Gene Smock is architecture.
Thought 5: The child of Gene Smock is Williams Smock. Now I need
    to retrieve article about Williams Smock and find his daughter
    , to get the great-granddaughter of Alvaro Smock.
Action 5: RetrieveArticle[Williams Smock].
Observation 5: # Williams Smock ## Family The parents of Williams
    Smock are Dominique Smock, Gene Smock.  The mother of Williams
     Smock is Dominique Smock.  The father of Williams Smock is
    Gene Smock.  The children of Williams Smock are Shelli Beltran
    , Stacia Toombs.  The daughters of Williams Smock are Shelli
    Beltran, Stacia Toombs.  The wife of Williams Smock is Alison
    Smock.  ## Friends The friend of Williams Smock is Lannie
    Smock.  ## Attributes The date of birth of Williams Smock is
    0926-04-04.  The occupation of Williams Smock is clinical
    biochemist.  The hobby of Williams Smock is social studies.
Thought 6: The daughters of Williams Smock are Shelli Beltran,
    Stacia Toombs. So the great-granddaughters of Alvaro Smock,
    whose hobby is biology, are Shelli Beltran, Stacia Toombs. The
     answer is Shelli Beltran, Stacia Toombs.
Action 6: Finish[Shelli Beltran{constants.answer_sep}Stacia Toombs
    ].
(END OF EXAMPLES)

Now answer the following question:
Question: {{question}}
{{scratchpad}}
```

D ADDITIONAL RESULTS